딥 러닝 기반 신호 검파기 모델 경량화 기술 연구

김연군, 이정우
서울대학교 전기정보공학부
ygoonkim@cml.snu.ac.kr, junglee@snu.ac.kr

Applying Model Compression Methods on Deep Learning based Symbol Detector

Yeongoon Kim, Jungwoo Lee
Department of Electrical and Computer Engineering, Seoul National University

요 약

최근 통신 채널을 통해 전달된 출력에서 노이즈를 제거하여 입력을 복원하기 위한 신호 검파기 기술에서, 선행 정보가 요구되는 채널 상태 정보(CSI)를 예측하기 위해 딥 러닝 기반 인공신경망이 많이 활용되고 있다. 선행 딥 러닝 기반 신호 검파기 연구들은 기존의 Viterbi, BCJR 알고리즘과 비슷한 성능을 달성하였지만, 이를 사용하기 위한 모델의 크기는 여전히 소형 IoT 기기 등에 적용하기에는 큰 실정이다. 본 논문에서는 딥 러닝에 사용되는 주요 인공신경망 경량화 기술인 양자화, 가지치기 기법을 실험을 통해 CSI 예측 인공신경망에 적용해보고, 신경망 정보 보존량에 따른 성능 보존 정도를 확인해본다.

Ⅰ. 서론

디지털 통신 시스템에서, 신호는 기본적으로 노이즈가 포함된 채널을 통해 송수신이 이루어진다. 따라서 출력은 이에 영향을 받아 입력에 에러가 추가된 형태로 전달받게 되고, 수신된 신호로부터 해당 에러를 제거하여 원래 신호를 복원하기 위한 신호 검파기(Symbol Detector) 기술을 필요로 한다. 신호를 복원하기 위해서는 신호 코드 단위 길이의 가변 여부, 채널 정보 유무, 채널 상태 및 노이즈의 변동성 등 다양한 요소를 고려하여 알고리즘을 적용해야 하며, 대표적으로 사용되는 알고리즘으로는 Viterbi[1], BCJR[2] 알고리즘 등이 있다.

위와 같은 model-based 알고리즘들은 채널 상태 정보(Channel State Information, CSI)가 충분히 주어졌을 때는 우수한 복원 성능을 보여주지만, 그렇지 않을 때는 예측이나 가정에 의존하기 때문에 보다 실제 환경에 가까운, CSI가 부정확하게 주어지거나. 가변성을 띄거나. 혹은 불분명한 상황에 적용이 어렵다. 이를 해결하기 위해, 최근에 급격히 발전하고 있는 딥 러닝(Deep Learning) 기법을 입력 신호 예측에 적용하는 방법이 대두되고 있다. 최근에 발표된 ViterbiNet[3], BCJRNet[4], MetaSSD[5] 등의 연구에서는 기본적으로 CSI 정보를 임의로 예측하는 부분을 입출력 데이터를 학습한 인공신경망(Neural Network)으로 대체하여, 보다 정확한 CSI 정보를 알고리즘에 제공할 수 있도록 한다.

딥 러닝 기법을 적용한 디코딩 알고리즘으로 보다 정확한 복원을 수행할 수 있지만, 이를 반도체 chip, IoT 디바이스 등 소형 하드웨어에 적용하기 위해서는 인공신경망의 경량화 기술의 적용이 필수적[6]이다. 대표적인 인공신경망 경량화 기술로는 신경망 양자화(Network Quantization), 가지치기(Pruning), 지식 증류(Knowledge Distillation) 기법 등이 있다. 본 연구에서는 위와 같은 다양한 인공신경망 경량화 기술을 딥 러닝 기반 디코딩 알고리즘에 적용하는 방안을 탐색해보고, Python 시뮬레이터 실험을 통해 디코딩 성능 및 모델 크기 축소 성능 등을 검증하는 것을 목표로 한다.

Ⅱ. 본론

딥 러닝 기반 코드 디코딩 기술 딥 러닝 기반 신호 검파기의 대표적인 기술로는 ViterbiNet, BCJRNet이 있다. Viterbi 알고리즘과 BCJR 알고리즘은 각각 최대 우도 추정(Maximum Likelihood, ML)과 최대 사후 확률 추정(Maximum a Posterior, MAP)에 기반하여, 출력 코드와 CSI로부터 계산된 finite-memory channel의 path cost 또는 factor graph가 주어질 때 가장 error rate가 적은 예상 입력 신호를 추정하는 알고리즘이다. 이 때 CSI가 요구되는 path cost 계산 및 factor graph 구축을 입출력 코드로 구성된 신호 데이터로 학습 가능한 DNN(Deep Neural Network)로 대체하여, CSI가 주어지지 않고서도 Viterbi, 혹은 BCJR 알고리즘을 적용할 수 있도록 하는 것이 ViterbiNet, BCJRNet의 핵심 아이디어이다. 또한 최근에는 가변하는 채널 환경에 대응하기 위해 기존 네트워크에 실시간으로 학습을 가능하게 하는 온라인 학습(Online Learning), 또는 가변 채널에 대응하는 방식을 학습하는 메타 학습(Meta Learning) 기법을 적용하는 방법 또한 연구되고 있다[5].

인공신경망 경량화 기술 인공신경망 경량화 기술은 네트워크 성능을 최대한 보존하면서, 네트워크의 파라미터 개수와 학습 시간을 최소화하는 것을 목표로 한다. 최근 연구되고 있는 대표적인 경량화 기술로는 신경망 양자화, 가지치기, 지식 증류 기법 등이 있다. 네트워크 양자화 기법이란 인공신경망을 구성하는 floating-point type 파라미터를 integer-point type으로 양자화하는 기술이며, 파라미터 당 메모리가 감소하여 신경망 크기 및 연산량의 감소를 달성할 수 있다. 또한 지식 증류 기법이란 상대적으로 높은 성능을 가질 것으로 기대되는 큰 네트워크 (Teacher network)로부터 추출되는 지식을 작은 목표 네트워크 (Student network)에 전달함으로써, 결과적으로 모델의 크기를 압축하는 효과를 가져오는

기법을 말한다. 마지막으로 네트워크 가지치기 기법이란 큰 네트워크에서 상대적으로 덜 중요하다고 여겨지는 weight value나 multiplication path를 제거함으로써 네트워크의 경량화를 이뤄내는 기법을 말한다. 본 연구에서는 이 중 네트워크 양자화 기술 및 가지치기 기법을 적용하여 네트워크의 파라미터 수 및 성능 보존도를 비교하여 네트워크의 경량화가 달성되었는지를 확인한다.

실험 방법 본 논문에서는 MetaSSD[5] 논문에서 공개된 코드를 기반으로 학습한 네트워크를 대조군으로 지정하였다. 대조군 네트워크는 6개의 선형 레이어(Linear Layer)가 쌓인 네트워크이며, 총 파라미터 수는 157,658개, GPU 메모리 사용량은 1109MiB이다. 실험 사항 중 네트워크 양자화의 경우 torch.quantization.quantize_dynamic 함수를 사용하여 float32 파라미터를 int8 파라미터로 양자화하였다. 이를 통해 네트워크 크기를 약 1/4 크기로 압축할 수 있다. 또한 네트워크 가지치기의 경우 torch.nn.utils.prune에 구현된 global_ unstructured(전체 네트워크에서 가지치기 수행), random_unstructured(네트워크의 특정 레이어에서 가지치기 수행, 이번 실험에서는 90300개의 파라미터를 가진 중앙에 위치한 가장 큰 레이어에서 가지치기 수행) 두 가지 함수를 사용, 가지치기 강도를 바꿔가며 성능을 측정, 비교해보았다. 실험 결과는 다음과 같다.

SNR / method	0	3	7	11	15
원본	12.1	7.0	3.3	1.9	1.4
양자화	48.8	48.6	48.5	48.4	48.3
전체 가지치기 (20%)	12.1	7.0	3.4	1.9	1.4
전체 가지치기 (50%)	12.1	7.0	3.4	1.9	1.4
전체 가지치기 (80%)	12.1	7.0	3.3	1.9	1.4
전체 가지치기 (90%)	11.8	6.8	3.3	2.0	1.5
전체 가지치기 (95%)	11.6	6.9	3.7	2.4	1.9
특정 가지치기 (20%)	11.8	6.9	3.7	2.3	1.8
특정 가지치기 (50%)	11.9	7.2	3.9	2.7	2.2
특정 가지치기 (80%)	12.5	8.4	5.4	4.1	3.5
특정 가지치기 (90%)	13.6	10.1	7.6	6.3	5.8
특정 가지치기 (95%)	13.6	10.1	7.6	6.3	5.8

(단위: 신호 예측 오류율(Signal Error Rate)(%))

[표 1] 인공신경망 경량화 실험 결과

실험결과 및 논의 양자화 기법 실험의 경우 제대로 학습되지 못하고 인공신경망이 제대로 된 역할을 수행하지 못함을 보여주었다. 이를 보완하기 위해 양자화 인지 학습(Quantization Aware Trainig)을 적용하여 네트워크를 새롭게 학습하는 방법을 제안할 수 있다. Purning 실험의 경우 약 8~90% 가량의 인공신경망 weight 파라미터를 0으로 masking해도 성능이 보존되어, 상대적으로 뛰어난 성능을 보여주었다. 다만 현재 pytorch 시뮬레이터 상으로는 weight memory를 제거하는 것이 아닌 value를 0으로 만드는 형태로 동작하기 때문에 실질적인 메모리 및 자원 이득이 없어, 적절한 하드웨어 개발이 후속되어야 한다. 마지막으로 지식 증류 기법의 경우, 대부분 매우 큰 모델로부터 작은 모델로 지식 주입이 이루어지는데, 현재 딥 러닝 기반 신호 검파기에 사용되고 있는 인공신경망은 간단한 선형 네트워크이기 때문에 사용이 부적절하여, 실험 사항에서 제외되었다.

Ⅲ. 결론

본 논문에서는 신호 검파기에서 신호를 복원하는데 필요한 CSI 정보를 인공신경망을 활용하여 예측하는 딥 러닝 기반 코드 디코딩 기술에서, 현존하는 네트워크의 크기를 감소시킬 수 있는 방법들을 제시하고 실험을 통해 그 성능들을 비교해보았다. 그 중 가지치기 기법을 사용했을 때 파라미터 수를 약 80% 가까이 감소시키더라도 낮은 SNR 환경에서 성능이 거의 보존됨을 실험을 통해 확인하였다.

ACKNOWLEDGMENT

This work is in part supported by Samsung Electronics Co., Ltd(Contract ID: MEM210728_0001), National Research Foundation of Korea (NRF, 2021R1A2C2014504), Institute of Information & communications Technology Planning & Evaluation (IITP, 2021-0-02068) grant funded by the Ministry of Science and ICT (MSIT), INMAC, and BK21-plus.

참 고 문 헌

[1] A. Viterbi. “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm”. IEEE Transactions on Information Theory, vol. 13, no. 2, pp. 260-269. 1967.

[2] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, “Optimal decoding of linear codes for minimizing symbol error rate,” IEEE Transactions on Information Theory, vol. 20, no. 2, pp. 284-287, 1974

[3] N. Shlezinger *et al.* "ViterbiNet: Symbol Detection Using a Deep Learning Based Viterbi Algorithm," 2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), pp. 1-5, 2019.

[4] N. Farsad *et al.* "Data-Driven Symbol Detection Via Model-Based Machine Learning," 2021 IEEE Statistical Signal Processing Workshop (SSP), pp. 571-575, 2021.

[5] M. J. Park *et al.* "MetaSSD: Meta-Learned Self-Supervised Detection," 2022 IEEE International Symposium on Information Theory (ISIT), pp. 480-485, 2022.